DOCUMENT RESUME

ED 337 494 TM 017 335

AUTHOR Hacker, Jacob; Hathaway, Walter
TITLE Toward Extended Assessment: The Big Picture.
PUB DATE Apr 91
NOTE 21p.; Paper presented at the Annual Conferences of the American Educational Research Association (Chicago, IL, April 3-7, 1991) and the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Cost Effectiveness; *Educational Assessment; Educational Trends; Elementary Secondary Education; Measurement Techniques; Standardized Tests; *Student Evaluation; *Testing Problems; Testing Programs; Test Reliability; Test Validity; *Thinking Skills; Trend Analysis
IDENTIFIERS *Authentic Assessment; *Performance Based Evaluation

ABSTRACT

Testing and assessment that are "more authentic" (performance-based or alternative) represent the most pressing issue in education today. Some of the major criticisms leveled at standardized testing are examined, and the advantages and disadvantages of more authentic assessment are reviewed. A general direction for integrating traditional and innovative forms of assessment is proposed. Among criticisms of current tests that have been identified by the National Commission on Testing and Public Policy, two of the biggest criticisms are that standardized objective testing fails to assess real mastery and is of limited validity. Advantages claimed for more authentic assessment include: (1) direct measurement of what children should know; (2) emphasis on higher order thinking skills, judgment, and collaboration; (3) encouragement of active participation in the learning process by children; and (4) allowing educators to teach to the test without destroying validity. Disadvantages of authentic assessment include high cost; difficulty in making results consistent and usable; and undemonstrated validity, reliability, and comparability. Three examples of authentic assessment are provided. A compromise between traditional and authentic assessments could begin with encouraging use of multiple measures and promoting more authentic measures when possible and cost effective. A 52-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

WALTER HATHAWAY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED337494

TOWARD EXTENDED ASSESSMENT: THE BIG PICTURE

Jacob Hacker and Walter Hathaway
Portland (Oregon) Public Schools

AERA/NCME/DRE/NATD Annual Conference

April 1991
Chicago, Illinois

TM017335

BEST COPY AVAILABLE

Testing and assessment that is "more authentic" or performance based or alternative is perhaps the most pressing issue in the profession of education today. Many questions are being asked about it. Is "more authentic" assessment being proposed for classroom use or as a substitute for traditional, large scale system accountability measures? What is an "authentic" task? What is the result of using far fewer, but more complete and in depth performance tasks than with traditional tests? How do we develop rubrics able to ensure consistency of response? How do we avoid bias, report error and ensure fairness? How do we ensure generalizability across tasks? What implicit theories of knowledge and student learning are being used? What are the underlying metacognitive and epistemological issues? What is reliability and validity in the context of "more authentic" assessment and how are they calculated? What are the effects on schools, classrooms and students of the greater time allocation required for the new assessment methods? How far ahead of the professionals are the politicians and how can we close the gap? i.e., How feasible and cost effective are "more authentic" assessments, especially for high stakes, large scale assessments?

The rapid rise of interest in "more authentic" assessment is causing a growing dilemma in the field of educational assessment. Two powerful trends seem to be on a collision course. On the one hand many policy makers are seeking and mandating increased educational reform and accountability, while on the other hand, many teachers, principals, administrators and theoretitions are turning increasingly toward school restructuring, teacher empowerment, and integrated curricular approaches as the vehicles to meaningful educational improvement. The source of the magnetic appeal of "more authentic" assessment for teachers and curriculum and instruction people may be that it promises both freedom from top down accountability and enhanced control and success in the classroom.

On the one hand the increasing press for educational accountability and productivity at all levels has led to a dramatic rise in system wide student assessment systems, most of them of a traditional, objective, standardized, multiple-choice, testing variety. For example, all fifty states now have some form of standardized, statewide assessment. On the other hand, there has been a move to restructuring for enhanced school autonomy and teacher control and toward more holistic approaches to curriculum and evaluation. These promise greater student learning, especially of thinking skills. Concerns such as these have caused increased resistance to traditional testing and growing interest in "more authentic" assessments.

The dilemma is, therefore, that policy makers who want to evaluate the success of systemwide educational reforms usually want more traditional testing while those educators committed to "grass roots" school and classroom educational reform and restructuring often prefer to forego or at least de-emphasize traditional tests in favor of "more authentic" assessments.

The phrase "more authentic" assessment means the gathering and evaluation of evidence of student performance produced in an integrated manner and in a naturalistic time frame and context. These assessments seek to reveal student performance on meaningful and challenging tasks as close as possible to the ones which the student is expected to perform in the "real world." There are several key characteristics of "authentic" assessment. It should: a) require a short chain of inference from the test performance to the real world competence (direct relevance); b) foster disciplined inquiry; c) challenge the student to integrate knowledge; and d) have value beyond evaluation (Archibald and Newman, 1988).

These assessments are often based upon performances, demonstrations, open ended questions, exhibitions, portfolios, or projects. For example, if a student is being assessed in science, he or she may be asked to perform a series of scientific experiments under expert observation and evaluation rather than take a traditional, multiple-choice, standardized, norm referenced science test. If the student is expected to write persuasive essays, then he or she will write persuasive essays in a naturalistic time frame and context; much as when a person is expected to ice skate or dive proficiently, he or she is asked to perform and is evaluated by expert observers.

At the same time as policy makers are demanding more traditional testing for evaluating the success of educational reform efforts, American students consistently perform at levels far below those achieved in the majority of industrialized nations (Shanker, pg. 1). A number of people believe that an over reliance on standardized testing itself may be a primary factor in America's educational lag. "The U.S. is the only nation that relies on multiple-choice tests for large-scale assessment," states Linda Darling-Hammond. "Most countries we compete with in Europe and Asia that out achieve us use essays, oral exams and exhibits of students' work" (Newsweek, Jan. 8, 1990). In response to growing concern over American students' poor international showing, members of the National Commission on Testing and Policy (1990) have recommended that alternative forms of assessment be adopted in American schools. More and more American educators, especially teachers and curriculum specialists, are demanding that testing become "more authentic", i.e., assess in a realistic and integral way meaningful skills and abilities including those of higher thinking and problem solving that enable students to become successful, productive adults. To the proponents of "more authentic" assessment and to the opponents of standardized testing, many of the evaluation tools currently used in America's schools provide little worthwhile information, lack "authenticity" and, ultimately, may undermine and subvert the educational process itself.

This paper: a) examines some of the major criticisms being leveled at standardized tests and misuses of their results; b) describes and discusses the claimed advantages and disadvantages of "more authentic" assessment; and c) proposes a general direction that might be taken toward integrating traditional and newer forms of assessment.

Criticisms of Current Standardized Testing

The National Commission on Testing and Public Policy (1990) has identified what it believes to be some key problems with standardized testing as it now exists:

1. Current tests are imperfect and therefore potentially misleading as measures of individual performance in education and employment.

2. Some tests result in unfair treatment of individuals and groups.

3. Students are subjected to too much testing in this nation's schools.

4. Some testing practices in both education and employment undermine important social policies and issues intended to develop or utilize human talent.

5. Tests have become instruments of public policy without sufficient public accountability (Commission Report, pg. 6).

Perhaps the biggest complaint leveled against standardized, objective achievement testing is that it fails to assess "real" mastery and therefore is of limited validity as an assessment of student learning. "A true test asks students to show what they know and can do, not to spout unrelated facts they have memorized the night before" (Horace, March 1990, pg. 1). Traditional testing has long been criticized for "neglecting the kind of competence expressed in "authentic", real life situations beyond school -- speaking, writing, reading, and solving mechanical, biological, or civic problems" (Archibald and Newman, 1988, pg. vi).

These charges of invalidity and irrelevance are typically based upon theoretical beliefs and assumptions rather than empirical studies. There is, on the other hand, a considerable body of empirical research results accumulated over the last 75 years or more which supports the advantages of objective, multiple-choice items, including the increasing capacity of well designed tests made up of such items to tap "complex thinking processes, reasoning, evaluation of arguments, and the application of knowledges to new situations." For example, "...Objective tests prove to be more valid predictors of the quality of essays written under proper conditions than do essay tests" (Anastasi, 1982 pg. 398, 399).

Much of the problem with current testing seems intimately linked to two key traditional testing assumptions, decomposability and decontextualization. These two assumptions underlie almost all current, traditional testing practices and are being increasingly challenged.

Early psychological theories were based on the assumption that thought was made up of a number of independent pieces of knowledge and that all skills could be broken down into smaller and more easily measurable components. Thus, if you wished to test whether a person was a skilled reader, you needed only determine whether they were able to perform the key subtasks that made up the skill of reading. This approach has been harshly criticized in recent years by proponents of holistic and integrated approaches to curriculum and instruction. They maintain that complex abilities cannot be defined solely by their components and that the whole is greater than the sum of its parts (e.g. Anderson, 1983). Thus, while there may indeed be a high correlation between eg. student scores on multiple-choice verbal tests and their ability to perform a skill such as writing, it is feared that at least some students will typically be misclassified as poor writers or as having incomplete verbal skills after multiple-choice testing, when in fact a "more authentic" and holistic assessment might have more accurately indicated that they were in fact competent.

The second major assumption apparent in almost all standardized, multiple-choice, achievement tests is that each component of a complex skill remains unaffected by the context in which it is used. In other words, if a student is able to perform decontextualized editing, a common element of standardized verbal tests, they will also be able to perform similar skills when editing their own work. However, studies have shown that there can be no absolute line drawn between data and its interpretation (e.g., Lakatos, 1978; Toulmin, 1972). The context in which a skill is performed is relevant; "knowledge and skill cannot be detached from their context of practice and use" (Resnick, pg. 9). Decontextualization of traditional test questions, then, is pointed to by critics as another factor that might hinder a traditional educational test in measuring accurately the broad abilities it purports to.

Also, according to the critics of traditional testing, the burden of the misclassifications they cause fall disproportionately on certain ethnic and linguistic minority groups as well as on students who have special learning needs, styles, or difficulties. The reasons posited for the observed disparity between minority and majority traditional test scores include cultural test bias, differences in economics or education, and the limited power of existing tests to predict success. Whatever causes such disparity, the fact remains that many minorities are currently being denied opportunities. Whenever testing limits the choices of individuals in certain groups, our assessment practices must be reexamined. On the other hand, it is the burden of any proposed replacement assessments to demonstrate their relative lack of bias and freedom from misclassification.

To the opponents of traditional, standardized testing the investment in it is also excessive. They note that over 20 million school days are used and 700 to 900 dollars annually for simply taking standardized tests. Even more importantly, it seems to them that tests are becoming more and more widely used for such controversial practices as kindergarten promotion and advancement from grade to grade, placement in "special learning" programs, and graduation from high school. Moreover:

* From 1972 to 1985 the numbers of state testing programs skyrocketed from 1 to 34. Every state now has a mandated testing program of some kind.

* Actual revenue from sales of tests and elected services has been estimated a half billion dollars per year.

* The direct cost for state and local testing plus indirect teacher costs may be as high as 915 million dollars annually (Commission Report, pg 7).

These figures, the critics say, fail to include another significant cost of so much standardized testing -- learning opportunity cost. Much of the time spent teaching the routine and lower-order thinking skills often present on standardized tests could be put to much better use. In an effort to improve increasingly "high stakes" test scores, many educators have resorted to spending inordinate amounts of class time actually "teaching to the test." Such huge fiscal and opportunity costs could be justified as legitimate educational expenses if they positively affected our school systems. However, the critics say the continuing trend of increased standardized testing has not created appreciable improvement in student performance (e.g., Shanker, pg. 1; Commission Report, pg. 18).

One of the primary functions of testing is to assess educational quality. In recent years local attention to reform has been directed by mandatory, "high stakes" testing. The danger in such testing is that when the stakes are raised, the pressure on schools to improve their scores may lead to disastrous solutions and undermines the educational process itself. In Pennsylvania, for example, mandatory state test scores were made public in 1987. Immediately the test scores became a "benchmark" for comparison between Pennsylvania school systems -- the tests became "high stakes." Schools that performed poorly lamented that they would have to alter their curriculum for the following year. One superintendent explained, "We don't believe in the test that strongly, but we will be forced to see that all material is covered before the tests...We won't be caught in the newspapers again" (Corbett and Wilson, 1989). Others involved in similar dilemmas agreed. "Teachers feel jerked around," a Maryland teacher confided. "The test dictates what I will teach in my classroom" (Corbett and Wilson, 1989). The charge of the opponents of standardized testing is, then, that more and more schools are becoming involved in "high stakes" testing and are therefore led to "teach to the test"

in order to raise test scores. Improving test results becomes more important than other, arguably more important, teaching and learning and societal responsibilities.

The final criticism of traditional testing rests on the apparently fatal allure of one point in time test scores in isolation to all other actual or possible kinds of evidence. Students are placed in Talented and Gifted Programs or remedial programs largely on their scores. Programs, policies, budgets and professionals all rise and fall with test scores. And this in the face of an almost universal commitment within education to using multiple indicators to support important educational decisions. When teachers spend valuable class time emphasizing test taking strategies, children learn those skill components and test strategies rather than higher level processes (N. Frederiksen, 1984). As Grant Wiggins, President of CLASS (Consultants for Learning Assessment and School Structure), points out, "What you test is what you get." "If we want to have quality assessment that creates quality work, we need to test for the task we want kids to be good at (from Videotape "Multi-dimensional Assessment Strategies," 1990). This point of view of the tyranny of test over teachers seems to regard them as something less than fully professional.

If "high-stakes" tests were adequate measures of students' performance, then they would serve to reinforce curriculum and aid learning. When current tests become too important, however, say their critics, school curriculum is actually debased because it focuses on simplistic multiple-choice questions and test-taking skills (Koretz, 1988).

There are a number of advantages claimed for "more authentic" alternative assessment techniques:

a) they measure directly what children should know;

b) they emphasize higher thinking skills, personal judgment and collaboration;

c) they urge children to become active participants in the learning process; and

d) they allow and encourage educators to "teach to the test" without destroying validity.

There are also, however, a number of possible disadvantages. These include:

a) high cost;

b) difficulty in making results consistently quantifiable, objective, standardized, and aggregatable; and

c) undemonstrated validity, reliability and comparability of the more subjective scoring systems and their results.

Advantages

One of the greatest advantages claimed for "authentic" testing is that it can test directly what educators want children to know. Because "authentic" testing assumes neither decomposability nor decontextualization, important skills can be tested "holistically" and in context. Holistic and high context testing lead to the mastery of the desired performance.

"Authentic" assessment also emphasizes the skills of higher thinking and personal judgement and allows collaboration. Performance tests can allow students to write, speak, listen, create, do original research, analyze, pose and solve problems, while much of standardized testing fails to even approximate such tasks according to some critics (e.g., Resnick 1989; Shanker, 1990). Peter Elbow and Pat Belanoff, examiners of a progressive writing program at the State University of New York at Stony Brook, discovered that "authentic" assessment teaches students "that their reactions and opinions about serious matters deserve time and attention," whereas standardized tests often stifle creativity and personal insight because the multiple-choice format implies that all the students can do is choose (or guess) someone else's "right" answer (Resnick, 1989). The proponents of traditional tests point to the value of recognition as well as generation both in school and in life. They also point to the ways in which skilled test designs use recognition of correct responses to get at varied, higher order skills. Such a format, however, does not allow the students to engage in direct interpretive activity and ultimately may leave the test-taker feeling powerless and uninvolved. ""Authentic" assessment, on the other hand, is designed to create an environment in which students can "show" what they know, leaving the power in their hands and allowing them to utilize higher thinking skills (Horace, March 1990).

"Authentic" assessment also helps children become more involved in their own learning process. Howard Gardner of Harvard's Project Zero claims that there are seven basic types of intelligence: linguistic, musical, spatial, logical/mathematical, bodily kinesthetic, interpersonal and intrapersonal. The majority of class time and standardized testing are focused on only two of these types: linguistic and logical/mathematical. Two very important types of intelligence, interpersonal and intrapersonal, are often neglected. Well developed intrapersonal intelligence is a common trait in successful individuals. Most "authentic" testing involves some form of self-criticism and personal evaluation, whether it be editing a piece of writing or critiquing a drawing. Most standardized testing, however, is thought to involve other peoples' work (editing someone else's writing, solving problems using predetermined techniques, etc.) and actually discourages interpersonal intelligence (Resnick, 1989; Archibald and Newman,

1989). Interpersonal intelligence, the ability to relate with others, is also claimed to be fostered with "authentic" assessment.

Many educators also feel that new forms of assessment should be collaborative (e.g., Valencia, McGinley and Pearson, in Press). In the world beyond school, students will usually have to work and create with others; rarely does someone in the "real world" create and perform without outside criticism and help. Collaborative assessment helps students develop their intrapersonal intelligence and strengthens the bond between teacher and student (e.g., Valencia, McGinley and Pearson, in Press; Elbow and Belanoff, 1986).

Arguably, the most important advantage of "authentic" assessment is that it allows tests to be instructional. Rather than be an after-the-fact check-up on students' learning, "authentic" tests can reinforce the curriculum and establish genuine intellectual standards. Thus, teachers can "teach to the test" without undermining the validity of the test. In fact, with "authentic" assessment, "teaching to the test" is not only possible, it is desirable (Resnick, 1989). Such an attitude conflicts with the general assumption that "teaching to the test" is a poor practice. However, with current standardized testing "teaching to the test" is indeed problematic, mainly because of the concept of **indicators**. While a high verbal score on the SAT may be an **indicator** of how well a student will perform on an actual written composition, the student need only be able to perform well on multiple-choice-type questions to **indicate** this ability. Such testing assumes that the student is being taught proper writing skills in the classroom. But, as pointed out earlier, students taking standardized tests need only to be able to perform specific test exercises to score well (Cannell, 1989). If "authentic" testing measures the skills and abilities educators believe are crucial in performing beyond school, then teaching to the test will raise school standards, improve curriculum and benefit society (Wiggins in Education Week, 1989). Thus, a major claimed advantage of "authentic" testing is that it frees educators from spending excessive time on a minimal, test directed curriculum. Again, there might be alternate routes to this liberation, such as regulations and rules preventing the use of student test scores to evaluate teachers (and principals).

Disadvantages

Although the costs of standardized testing today are staggering, "authentic" assessment could prove to be many times more expensive. The need for increased professional time for assessment and such costly items as video cameras could increase assessment costs significantly. For example, in the R.O.P.E. (Right Of Passage Experience) program used at Walden III, an alternative public school in Racine, Wisconsin, at least 10 hours of extra teacher time are needed for **each** graduating student (Horace, March 1990). In a school with a graduating class of 500, that would amount to at least 5,000 more paid hours per year! At 20 dollars an hour, teaching costs alone would increase $100,000 per

school per year! While there are many different estimates of potential cost, it is clear that "authentic" assessment requires far more teacher and student time than computer-scored multiple-choice tests or even than emerging versions of traditional tests adapted to include some open-ended items and other changes to respond to new curriculum and instructional directions. Because few states, districts or schools have utilized extensive amounts of "authentic" assessment, except in pilot versions, actual costs remain unclear.

Further difficulties in "authentic" assessments stem from the problems encountered in attempts to make their results valid, reliable and comparable. Here the key issue is subjectivity in evaluating performances. It is difficult to assign a specific, adequately discriminating, scaled score or percentile to a "more authentic" assessment, such as an essay, and it is even harder in the case of portfolio evaluations, etc. compared to having a computer count the number of wrong responses to the items on a well designed, objective, standardized test. Rarely does a scale on a performance-based assessment contain more than 10 points. In research done by the Portland Public Schools, we have demonstrated that the results of Direct Writing assessment done in a manner consistent with writing as a process are invariably prompt dependent. As a result, systemwide assessment of writing performance cannot be validly compared over time. Thus, it is not possible to answer the question, "Is this eighth grade doing better or worse than last year's?" But this is just the sort of question policy makers need and want to have answers to so that they can modify policies, programs and resources in productive ways. "Authentic" assessments alone thus far cannot readily serve all the decision making needs of educational policy makers, planners, designers and resource allocators beyond the individual classroom. A sense of the nature and strengths and weaknesses of "more authentic" assessments can be gained by investigating some examples of applications such as follow.

Examples

Walden III's R.O.P.E. Program

Walden III, an alternative public school in Racine, Wisconsin, has developed a program to address the issue of student preparation for life beyond school. In order to graduate each senior must demonstrate mastery in 15 areas of knowledge and competence by completing and submitting a portfolio of work before a committee made up of staff members, another student in a lower grade, and an adult from the community. The portfolio includes: an autobiography, self-analysis, essays, artistic products, letters of recommendation, and various other indicators of mastery. The portfolio itself is presented by the student before the committee and carefully evaluated and approved before graduation can occur. Clearly, Walden III's program meets the first three criteria, direct relevance, disciplined inquiry and integration of knowledge admirably. In addition, the forth characteristic, value beyond evaluation, is fulfilled by the actual

process of completing the portfolio itself. The student may spend more than two years working on the project outside of class. They have as long as they like beginning in their junior year. All the time spent is both educational and self-directed, allowing the student to learn the responsibility and self-discipline which will be needed in college and in later life.

Key School, Indianapolis, Indiana

The Key School is the child of Professor Howard Gardner of Harvard's Project Zero. Located in Indianapolis, Indiana, the fifth grade school is one of the most progressive public schools in the nation. The school utilizes video cameras to tape all projects and oral tests that the students' complete. A full time video technician helps keep a video file on each child which can be viewed by students, teachers and parents alike. The classroom environment is non-competitive and the school's philosophy is to build students' strengths rather than reinforce weaknesses. From all accounts, the school has been very successful. However, the cost of all the equipment and extra teaching time is very high. The Key School is experimental and there are few similar programs in existence for financial and logistical reasons. Clearly, it too meets all four of the "authentic" assessment criteria.

Michigan Educational Assessment Program (MEAP) - An Accommodation

The Michigan Educational Assessment Program (MEAP) was established in the late 1960's to provide information on the progress of Michigan students in the essential skills areas. However, when it was decided that these tests no longer provided Michigan educators with adequate feedback on the progress and status of Michigan basic skills education, a group of teachers and curriculum specialists designed the Michigan Essential Skills Reading Test. While the tests use a multiple-choice format, they are untimed. The test also attempts to measure attitudes about reading and self-perceptions of the test-takers. The passages read are long (e.g., 500-2,000 words) and the questions, although multiple-choice, are designed to challenge the reader to construct meaning from the text. In addition, the test is designed to assess the familiarity the test-taker has with the reading selection topic. In context, this contradicts the theory that reading assessment selections should be interest and curriculum neutral and context free. Instead the test assesses the student's relevant prior knowledge and experience. These characteristics allow the MEAP test to meet the first three criteria outlined earlier -- disciplined inquiry and integration of knowledge -- the final and arguably most important criteria, value beyond evaluation, was tackled by the Michigan program also. Test result forms were designed in such a manner that the student, teacher and parent can immediately see not only the student's performance in individual areas but also the influence of each performance on other areas. For example, if topic familiarity is low, then lower scores on other sections might be a result of inadequate knowledge of the

topic. If the self-perception section indicates that the child is uninterested, then the teacher or parent can immediately try to bring their interest level up. The Michigan Program is promising because it shows the degree and limits to which the tenets of "more authentic" assessment can be accommodated within the less resource-intensive and well-established standardized, multiple-choice format.

Some Possible Steps Forward

A possible accommodation between alternative, "more authentic" assessment and traditional standardized measures of academic achievement could go as follows:

1. Accept, promote and practice the belief that multiple measures are better than single ones, especially in measuring gain from one assessment to another versus one time levels of performance;

2. Embrace the desirability of "more authentic" measures whenever possible and cost effective, for example, in sampling approaches to large system data gathering;

3. Encourage and support development and use of "more authentic" assessment techniques by teachers in their classrooms for assessing and monitoring their students' progress and their needs for further learning opportunities and experiences;

4. Work with curriculum and instructional professionals to modify standardized, multiple-choice testing systems so that they:

   A. Assess new dimensions such as context and prior knowledge.

   B. Develop and add to standardized assessment systems "more authentic" items, eg. for Reading, obtain permission to use long passages of connected, meaningful text from "published" materials and ask multiple questions, at least some of which tap higher order thinking skills; or for Mathematics, use everyday problems and permit use of

calculators for all items except those which assess computation and estimation, etc.

C. Add open ended, extended, non-multiple-choice items to standardized multiple-choice tests. For example, on a Mathematics test pose problems and give students time and space to work out the answers and a place on the answer sheet to code in their responses. Use information from such adaptations in an integrated fashion along with the traditional scores.

D. Add additional "more authentic" items in system- wide assessments using a (matrix) sampling design in order to lend depth of insight into the meaning of large group, aggregate data while maintaining cost effectiveness.

5. Perform the necessary research, development and evaluation collaboratively among R & D centers, school systems and university to:

A. Test the boundaries of the construct and predictive validity and the reliability of "more authentic" assessment techniques, particularly in large-scale assessment environments.

B. Develop the computer/video disk system and software necessary to extend the power of current computer adaptive testing systems so that the new systems simultaneously provide personalized, cost-effective instruction environment as well as continuous, very valid, objective, highly realistic and "authentic" assessment of student learning.

C. For example, a state assessment program designed along the lines sketched above could consist of the following:

1. Well designed mostly standardized, multiple-choice tests both NR and CR designed to sample domains of the ...state curricula.

2. Selected open ended items, surveys of prior knowledge, and attitudes practices, etc. as the subject dictates.

3. Alternative performance assessments of the type being encouraged and supported in classrooms administered selectively as resources permit.

Such an approach would provide both complex varied, accurate data for state as well as local decision making. It would also model and reinforce the use of performance assessments at the school and classroom level with state and local support.

Thus for long range solutions to our dilemma, we must turn once again to research and development and to evaluation. Here we need to continue, extend and evaluate pilot efforts to develop, use in a cost effective way, and test the validity reliability and efficacy of such non-traditional assessment approaches as portfolios, projects and performance assessments. In doing this we will need to pay attention to the problems of the consistency over different raters and over time of rating-like responses we need to gain accurate information with the new methods and their scoring systems on scales of sufficient range to permit meaningful and necessary discriminations. We must also pay particular attention to the need for developing and reporting accurate portrayals of the degree of error and uncertainty of the estimates yielded by "more authentic" assessments. We need to research the construct, content, predictive-criterion and face validity, as well as the relevance of each type of measure and of reports of their results. At the same time we must continue to work to adapt and evaluate traditional measures to make them "more authentic" and to experiment with the ways in which technology can help us.

Another useful area to continue and expand psychometric research and development is computerized (adaptive) testing, especially cutting edge systems in the areas of artificial intelligence, expert systems, fuzzy logic and video disk/computer interfaces. These systems have the promise in the long run of assessing the higher order thinking and problem solving skills that critics of traditional standardized testing say it fails to adequately assess.

None of these recommendations call for an immediate, wholesale overhaul of existing testing procedures. Instead they urge immediate implementation and evaluation of the emerging assessment methodology in those relatively low stakes areas (such as classroom assessment) and continued and extended research and development for implementation in relatively high stakes areas (such as system wide assessment for evaluation accountability and planning and graduation competence certification). Since assessment's fundamental purpose should remain as helping children, their parents and their teachers receive useful feedback, both positive and negative, about the performance and needs of students and the education system which serves them. The controversy over standardized vs. "authentic" assessments unfortunately diverts attention from the more important mission. Tests and assessments are only tools; they can be either valuable or worthless depending on where, when and how they are used. The solution to our current dilemma is not as simple as saying "no more standardized tests." Perhaps we should be saying "no more closed doors" and "no more closed minds."

Bibliography

Anastasi, A. (1982). Psychological testing (5th ed.). New York, NY: Macmillan Publishing Co., Inc., London, England: Collier Macmillan Publishers.

Archibald D.A., and Newman, F.M. (1988). Beyond standardized testing: Assessing authentic academic achievement in the secondary school. Reston, VA: National Association of Secondary School Principals. Anderson, J.R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Ascher, C. (1988, August). Grade retention: Making the decision. ERIC/CUE Digest No. 46. New York: ERIC Clearinghouse on Urban Education, Teachers College, Columbia University.

Baker, E.L. (1989). Mandated tests: Educational reform or quality indicator? In B.R. Gifford (Ed.), Test policy and test performance: Education, language and culture. Boston: Kluwer Academic Publishers.

Belanoff, P., and Elbow, P. (1986). Using portfolios to increase collaboration and community in a writing program. In Cassally and Villard (1986). New methods and college writing programs. MLA.

Bond, L. (1988, December). Understanding the black/white student gap on measures of quantitative reasoning. Presented at Testing and the Allocation of Opportunities in Education and Employment to Black Americans, Howard University, Washington, DC. Unpublished manuscript. University of North Carolina.

Brandt, R. (1989). On misuse of testing: A conversation with George Madaus. Washington, D.C.: Educational Leadership.

Carnegie Forum on Education and the Economy. (1986). A nation prepared: Teachers for the twenty-first century. New York: Author.

Cannell, J.J. (1989). How public educators cheat on standardized tests. Albuquerque, NM: Friends for Education.

Catterall, J.S. (1987). Towards researching the connections between tests required for high school graduation and the inclination to drop out of school. Los Angeles: California University, Center for the Study of Evaluation.

College Board (1988). National college-bound seniors: 1988 profile. Profiles of SAT and achievement test takers national ethnic sex profile. New York: The College Board.

Corbett, H.D., and Wilson, B. (1989). Raising the stakes in statewide mandatory minimum competency testing. Philadelphia, PA: Research For Better Schools.

End of standardized tests requested. (1989, October 29). The Columbian.

Costa, A.L. (1989). Re-assessing assessment. Washington, D.C.: Educational Leadership.

Darling-Hammond, L., & Wise, A.E. (1985). Beyond standardization: State standards and school improvement. The Elementary School Journal, 3, 133-143.

Door-Bremme, D.W., & Herman, J.L. (1986). Assessing student achievement: A profile of classroom practices. Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation.

Gardner, H. (in press). Assessment in context: The alternative to standardized testing. In B. Gifford & M.C. O'Connor (Eds.), Future assessments: Changing views of aptitude, achievement, and instruction. Boston: Kluwer Academic Publishers.

Haertel, E. (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B.R. Gifford (Ed.), Test policy and performance: Education, language and culture. Boston: Kluwer Academic Publishers.

Hansen, J.B., Hathaway, W.E. (1988). A survey of more authentic assessment practices. Chicago, IL: A Presentation in the NCME/NATD Symposium.

Haney, W.M. (1989). Making sense of school testing. In B.R. Gifford (Ed.), Test policy and test performance: Education, language and culture. Boston: Kluwer Academic Publishers.

The Coalition of Essential Schools. (1990, March). Performance and exhibitions: The demonstration of mastery. Horace.

Kirst, M.W. (1991). Interview on assessment issues with Lorrie Shephard. Interview on assessment issues with James Popham. Washington, D.C.: Educational Researcher.

Koretz, D. (1988). Arriving in Lake Wobegon: Are standardized achievement tests exaggerating advancement and distorting instruction? American Educator. 12, 46-54.

Lakatos, I. (1978). The methodology of scientific research programs, philosophical papers. Vol. 1. J. Worrall & J. Currie (Eds.). New York: Cambridge University Press.

Lewis, M., Lindaman, A.D. (1989). How do we evaluate student writing? One district's ANSWER. Washington, D.C.: Educational Leadership.

Martinez, M.E., Lipson, J.I. (1989). Assessment for learning. Washington, D.C.: Educational Leadership.

Madaus, G., & Haney, W. (in press). The fractured marketplace for standardized testing. Boston: Kluwer Academic Publishers.

Michigan State Board of Education. (1989). Essential skills reading test blueprint. Lansing, MI. Unpublished.

Multidimensional assessment: Strategies for the classroom. Video tape #4 in the series Restructuring to Promote Learning in America's Schools. (1990).

National Alliance of Business. (1987). The fourth R: Workforce readiness. Washington, DC: Author.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: U.S. Government Printing Office.

National Commission on Testing and Public Policy. (1990). From gatekeeper to gateway: Transforming testing in America. Chestnut Hill, MA: Author.

Newell A., and Simon, H.A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice Hall.

Not as easy as A, B or C. (1990, January 8). Newsweek.

Nickerson, R.S. (1989). New directions in educational assessment. Washington, D.C.: Educational Researcher.

Oakes, J. (1987). Improving inner city schools: Current directions in urban district reform. Santa Monica: RAND Corporation, Center for Policy Research in Education.

Raizen, S.A., Baron, J.B., Champagne, A.B., Haertel, E., Mullis, I.V.S., & Oakes, J. (1989). Assessment in elementary school science education. Andover, MA: National Center for Improving Science Education, The Network, Inc.

Reich, R.B. (1988). Education and the next economy. Washington, D.C.: National Education Association, Professional and Organizational Development/Research Division.

Resnick, C.B. (1989). Tests as standards of achievement in schools. Paper prepared for the Educational Testing Service Conference. The Uses of Standardized Tests in American Education, New York.

Robinson, S.P. (1989). The agenda for reform in the use of standardized tests: Achieving the ideal of inclusiveness. Princeton, NJ: Educational Testing Service.

Roeber, E., Dutcher, P. (1989). Michigan's innovative assessment of reading. Washington, D.C.: Educational Leadership.

Rogers, V. (1989, May). Assessing the curriculum experienced by children. Phi Delta Kappan, 70 (9), 714-718.

Shanker, A. (1990). The social and educational dilemmas of test use. New York: Educational Testing Service.

Shepard, L. (1989, October). What is test misuse: Perspectives of a measurement expert. The users of standardized tests in American education. Speech presented at the Invitational Conference of the Educational Testing Service, New York.

Shepard, L.A. (1989). Why we need better assessments. Washington, D.C.: Educational Leadership.

Sternberg, R.J. (in press). CAT: A program of comprehensive abilities testing. In B. Gifford & M.C. O'Connor (Eds.), Future assessments: Changing views of aptitude, achievement, and instruction. Boston: Kluwer Academic Publishers.

Stiggins, R.J. (1987). Design and development of performance assessments. In I.T.E.M.S. (Fall 1987).

Toulmin, S.E. (1972). Human understanding. Princeton, NJ: Princeton University Press.

Valencia, S.W., Pearson, P.D., Peters, C.W., Wixson, K.K. (1989). Theory and practice in statewide reading assessment: Closing the gap. Washington, D.C.: Educational Leadership.

Vickery, T.R. (1988, February). Learning from an outcomes-driven school district. Educational Leadership, 45 (5), 52-56.

Wolf, D.P. (1987, December/1988, January). Opening up assessment. Educational Leadership, 45, (4), 24-29.

Wolf, D.P. (1989). Portfolio assessment: Sampling student work. Washington, D.C.: Educational Leadership.

7/23/91